J. Cryptology (1994) 7: 33–51

Journal of
CRYPTOLOGY

The Vulnerability of Geometric Sequences Based on Fields of Odd Characteristic*

Andrew Klapper
Department of Computer Science, University of Manitoba,
Winnipeg, Manitoba, Canada R3T 2N2

Communicated by Joan Feigenbaum

Received 3 March 1992 and revised 28 September 1992

Abstract. A new method of cryptologic attack on binary sequences is given, using their linear complexities relative to odd prime numbers. We show that, relative to a particular prime number p, the linear complexity of a binary geometric sequence is low. It is also shown that the prime p can be determined with high probability by a randomized algorithm if a number of bits much smaller than the linear complexity is known. This determination is made by exploiting the imbalance in the number of zeros and ones in the sequences in question, and uses a new statistical measure, the partial imbalance.

Key words. Cryptanalysis, Binary sequence, Geometric sequence, Linear complexity, Galois field.

1. Introduction

In several applications in modern communication systems, periodic binary sequences are employed that must be difficult for an adversary to determine when a short partial sequence (that is, a subset consisting of consecutive elements of the sequence) is known, and must be easy to generate given a secret key. This is true both in stream cipher systems, in which the binary sequence is used as a pseudo-one-time-pad [16], and in secure spread spectrum systems, in which the sequence is used to spread a signal over a large range of frequencies [14]. While theorists have long argued that such security can only be achieved by sequences satisfying a very general statistical test such as Yao's [17] and Blum and Micali's [2] next bit test, practitioners are often satisfied to find sequences that have large linear

* This project was sponsored by the National Security Agency under Grant No. MDA904-91-H-0012. The United States Government is authorized to reproduce and distribute reprints notwithstanding any copyright notation hereon.

complexities, thus ensuring resistance to attack by the Berlekamp–Massey algorithm [12]. Linear feedback shift registers are devices that can easily generate sequences with exponentially larger period than the size of their seeds [5], though with small linear complexity. Thus, much effort has gone into finding ways of modifying linear feedback shift registers, typically by adding some nonlinearity, so that the sequences they generate have large linear complexities.

The purpose of this paper is twofold. First, to argue that high linear complexity is inadequate for security, even when the only concern is with attacks using the Berlekamp–Massey algorithm. To this end, we exhibit binary sequences, i.e., sequences of elements of the field $GF(2)$, with high linear complexity, but which have low linear complexity when considered as sequences (whose elements happen to be only zero or one) over a larger finite field. Our second purpose is to demonstrate that a particular class of sequences, geometric sequences based on m-sequences over fields of odd characteristic, are insecure for stream cipher systems. We do so by first proving that, for a particular prime p, these sequences have low linear complexity when considered as sequences over $GF(p)$, and then giving an algorithm which inputs a relatively small partial sequence and determines p with high probability (the probabilities are over the set of partial sequences of a fixed size). This is made possible by the fact that these sequences are imbalanced.

In Section 2 we recall the definitions of linear complexity and geometric sequences, and some of the basic results concerning them. In Section 3 we derive an upper bound for the linear complexity of the geometric sequences in question. In Section 4 we exhibit an algorithm for finding the characteristic of the field of definition for th- m-sequence on which a geometric sequence is based. This algorithm is based on computing the partial imbalance of a sequence (the imbalance of a partial sequence). Its success depends on bounding the variance of the partial imbalance. This is a similar approach to one taken in an earlier paper by the author and Mark Goresky based on the partial period autocorrelation [9], but requires far fewer bits for success.

2. Definitions

A sequence $\mathbf{S} = (\mathbf{S}_i)$ over a field F is *linearly recurrent* if it satisfies a recurrence of length n,

$$\sum_{i=0}^{n} c_i \mathbf{S}_{k-i} = 0, \tag{1}$$

with $c_i \in F$ not all zero. The *linear complexity* of $\mathbf{S}$ over F is the smallest n such that $\mathbf{S}$ satisfies a linear recurrence of length n. Equivalently, the linear complexity of $\mathbf{S}$ is the length of the smallest linear feedback shift register over F which generates $\mathbf{S}$. We denote the linear complexity of $\mathbf{S}$ over F by $\lambda_F(\mathbf{S})$. It is well known [12] that $2\lambda_F(\mathbf{S})$ consecutive elements of $\mathbf{S}$ suffice to determine the smallest linear recurrence satisfied by $\mathbf{S}$ (or, equivalently, to synthesize a linear feedback shift register which generates $\mathbf{S}$). Moreover, the algorithm for this determination is highly efficient. Thus, for purposes of security, sequences must be found which have large linear complexity.

Now consider a binary sequence S. While it is perhaps most natural to think of S as a sequence of elements of $GF(2)$, and thus concern ourselves only with the linear complexity over $GF(2)$, we can in fact think of S as a sequence over any prime field $GF(p)$. When S is a binary sequence, we denote the linear complexity of S over $GF(p)$ by $\lambda_p(\mathrm{S})$. It is our purpose to show that, despite the fact that $\lambda_2(\mathrm{S})$ is large, $\lambda_p(\mathrm{S})$ may be small for some computable p. This is not the case for all sequences. For example, the sequence $0^{n-1}10^{n-1}1\cdots$ has $\lambda_p(\mathrm{S}) = n$ for all p. Of course this sequence has other undesirable properties such as a lack of balance.

The sequences we consider are known as *geomtric sequences* and have been considered by several authors with respect to their linear complexity and correlation properties [3], [4], [6], [7], [8], [18]. In particular, Chan and Games showed that, for q odd, geometric sequences can have high linear complexities (over $GF(2)$) [4]. For this reason, these sequences are candidates for applications that require easily generated sequences that resist chosen plaintext attacks. Geometric sequences are based on algebra over finite fields, and we recall first some of the basic concepts. See Lidl and Niederreiter's [10] or McEliece's book [13] for a more detailed treatment of finite fields.

Let p be a prime integer, let $q = p^e$ be a fixed power of p, and let $GF(q)$ denote the Galois field with q elements. For any $n \geq 1$, we denote the *trace function* from $GF(q^n)$ to $GF(q)$ by $Tr_q^{q^n}$, defined by

$$Tr_q^{q^n}(x) = \sum_{i=0}^{n-1} x^{q^i}.$$

$Tr_q^{q^n}$ is a $GF(q)$-linear function, and every $GF(q)$-linear function f from $GF(q^n)$ to $GF(q)$ can be written in the form $f(x) = Tr_q^{q^n}(Ax)$ for some $A \in GF(q^n)$. For any $m \geq 1$, $Tr_q^{q^{nm}}(x) = Tr_q^{q^n}(Tr_{q^n}^{q^{nm}}(x))$.

Let α be a primitive element of $GF(q^n)$, that is, $GF(q^n)$ consists of zero and the powers of α. The infinite periodic sequence U whose ith term is $\mathrm{U}_i = Tr_q^{q^n}(\alpha^i)$ is known as an *m-sequence* over $GF(q)$ of span n [10]. More generally, we can consider the sequence whose ith term is $Tr_q^{q^n}(A\alpha^i)$ for some fixed element A of $GF(q^n)$. This amounts to a cyclic shift of the sequence U, so we do not consider it to be a distinct sequence here. It is well known that every m-sequence of span n can be generated by a linear feedback shift register of length n over $GF(q)$. It has period $q^n - 1$, the maximum possible period for a sequence generated by a linear feedback shift register of length n over $GF(q)$. Moreover, every such maximal period sequence is (a shift of) an m-sequence [10, pp. 394–410].

Definition 2.1. Let n be a positive integer and let α be a primitive element of $GF(q^n)$. Let g be a (possibly nonlinear) function from $GF(q)$ to $GF(2)$. The binary sequence S whose ith term is

$$\mathrm{S}_i = g(Tr_q^{q^n}(\alpha^i))$$

is called the geometric sequence based on the primitive element α and feedforward function g.

Note that if p is odd, then in general a function from a field of odd characteristic to a field of characteristic two cannot be defined algebraically. We can, however, think of g as simply dividing $GF(q)$ into two subsets (or as the characteristic function of a subset of $GF(q)$. Alternatively, we can think of g as a function into $\{0, 1\} \subseteq GF(r)$ for any prime integer r. If $r = p$, then $g(x)$ can be expressed as a polynomial in x with coefficients in $GF(q)$.

The geometric sequence **S** is easy to generate if the feedforward function g is easy to compute. This is always the case, for example, if q is small. Such a geometric sequence is a binary sequence of period dividing $q^n - 1$. Geometric sequences with q odd have been used in applications where easily generated sequences with large linear complexities are needed, due to the following theorem of Chan and Games [4], in which we let $v = (q^n - 1)/(q - 1)$.

Theorem 2.2 (Chan and Games). *Let* **S** *be a geometric sequence,* $\mathbf{S}_i = g(Tr_q^{q^n}(\alpha^i))$, *and suppose* $g(0) = 0$. *Let* **T** *be the sequence* $\mathbf{T}_i = g(\alpha^{iv})$ *(note that* α^v *is a primitive element of* $GF(q)$*). Then*

$$\lambda_2(\mathbf{S}) = v\lambda_2(\mathbf{T}).$$

Thus, if g is chosen to maximize $\lambda_2(\mathbf{T})$ at $q - 1$, then $\lambda_2(\mathbf{S})$ will be $q^n - 1$, i.e., maximal. It is straightforward to generalize this theorem to the linear complexity λ_r over an arbitrary prime number $r \neq p$. The details are left to the reader.

Theorem 2.3. *Let* **S** *be a geometric sequence,* $\mathbf{S}_i = g(Tr_q^{q^n}(\alpha^i))$, *and suppose* $g(0) = 0$. *Let* **T** *be the sequence* $\mathbf{T}_i = g(\alpha^{iv})$. *Let* r *be a prime number,* $r \neq p$. *Then*

$$\lambda_r(\mathbf{S}) = v\lambda_r(\mathbf{T}).$$

In particular, as long as $g(x) \neq 0$ for some $x \in GF(q)$, then $\lambda_r(\mathbf{S}) \geq (q^n - 1)/(q - 1)$ for all primes $r \neq p$. Thus the only prime over which **S** can have small linear complexity is p, and we show in the next section that this is indeed the case.

3. Bounding λ_p

In this section we show that if **S** a geometric sequence based on an m-sequence over a finite field of characteristic p, then $\lambda_p(\mathbf{S})$ is far from maximal. Our results follow from the work by Zierler and Mills [18], Herlestam [6], [7], and Brynielsson [3] on the linear complexity of algebraic combinations of periodic sequences over fields. What amounts to a special case of these results was proved by MacWilliams and Mann [11] and Smith [15] in the context of measuring the rank of the incidence matrix of points and hyperplanes in a finite geometry.

3.1. *Preliminaries*

If F is a field, then the set of infinite sequences over F is an algebra, with addition and multiplication defined componentwise. The set of polynomials in one variable t with coefficients in F acts on infinite sequences over F, with t acting as the shift operator:

$$t(\mathbf{S}_0, \mathbf{S}_1, \mathbf{S}_2, \ldots) = (\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3, \ldots).$$

Moreover, this is an F-linear action of the algebra of polynomials with coefficients in F on the vector space of infinite sequences over F. The condition in (1) is equivalent to $f(t)\mathbf{S} = 0$, where $f(t) = \sum_{i=0}^{n} c_i t^i$. If $f_{\mathbf{S}}(t)$ is the smallest degree nonzero polynomial such that $f_{\mathbf{S}}(t)\mathbf{S} = 0$, and $g(t)$ is another polynomial such that $g(t)\mathbf{S} = 0$, then $f_{\mathbf{S}}(t)$ divides $g(t)$. The polynomial $f_{\mathbf{S}}$ is known variously as the connection or feedback polynomial of $\mathbf{S}$. Its degree is $\lambda_F(\mathbf{S})$.

The following is a consequence of the work of Zierler and Mills [18]. If f and g are polynomials over F, let $f \vee g$ be the monic polynomial whose roots are the distinct elements of the form $\gamma\delta$ where γ is a root of f and δ is a root of g (over an algebraic closure of F). By Galois theory, $f \vee g$ is defined over F.

Proposition 3.1. *Let* $\mathbf{S}$ *and* $\mathbf{T}$ *be linearly recurrent sequences over* F. *Then*

(1) $f_{\mathbf{S}+\mathbf{T}}$ *divides the l.c.m. of* $f_{\mathbf{S}}$ *and* $f_{\mathbf{T}}$, *and* $\lambda_F(\mathbf{S} + \mathbf{T}) \le \lambda_F(\mathbf{S}) + \lambda_F(\mathbf{T})$. *If* $f_{\mathbf{S}}$ *and* $f_{\mathbf{T}}$ *have no common roots, then* $f_{\mathbf{S}+\mathbf{T}} = f_{\mathbf{S}} f_{\mathbf{T}}$ *and* $\lambda_F(\mathbf{S} + \mathbf{T}) = \lambda_F(\mathbf{S}) + \lambda_F(\mathbf{T})$.

(2) $f_{\mathbf{ST}}$ *divides* $f_{\mathbf{S}} \vee f_{\mathbf{T}}$ *and* $\lambda_F(\mathbf{ST}) \le \lambda_F(\mathbf{S})\lambda_F(\mathbf{T})$ = *the number of distinct root products* $\gamma\delta$, γ *a root of* $f_{\mathbf{S}}$, δ *a root of* $f_{\mathbf{T}}$. *If all the root products from* $f_{\mathbf{S}}$ *and* $f_{\mathbf{T}}$ *are distinct, then* $f_{\mathbf{ST}} = f_{\mathbf{S}} \vee f_{\mathbf{T}}$ *and* $\lambda_F(\mathbf{ST}) = \lambda_F(\mathbf{S})\lambda_F(\mathbf{T})$.

We also need the following well-known lemma.

Lemma 3.2. *The distinct roots of* $f_{\mathbf{S}}$ *are* $\{\gamma_j, j = 1, \ldots, n\}$ *if and only if* $\mathbf{S}$ *can be written uniquely as* $\mathbf{S}_i = \sum_{j=1}^{n} c_j \gamma_j^i$. *(The* γ_j *may lie in an extension field of* F. *The* c_j *will lie in* $F[\gamma_1, \ldots, \gamma_n]$.*)*

3.2. *Geometric Sequences*

In our situation we have an m-sequence $\mathbf{U}$ of period $q^n - 1$ over $GF(q)$, hence of linear complexity n over $GF(q)$. To this sequence we apply a feedforward function g: $GF(q) \to GF(q)$. For now g is arbitrary, but we eventually specialize to the case where the image of g is in $\{0, 1\}$. We can express g as a polynomial

$$g(x) = \sum_{i=0}^{p^e-1} a_i x^i$$

with $a_i \in GF(q)$ (recall $q = p^e$). The resulting sequence $\mathbf{S}$ can be thought of as a sum of constant multiples of powers of the original sequence $\mathbf{U}$. Herlestam [7] derived a formula for the linear complexity of such sequences in the case where g contains a single term. He also gave formulas relating the linear complexities of sequences to the linear complexities of their sums and products. However, he stopped short of considering a fully general feedforward function g applied to an m-sequence, and this is our situation. We use the results of the preceding subsection to express the linear complexity of $\mathbf{S}$ in terms of p, e, n, and $\{a_i\}$. Note that the connection polynomial $f_{\mathbf{U}}$ of $\mathbf{U}$ is the minimal polynomial of α over $GF(q)$, and α is a primitive element of $GF(q^n)$. Hence $f_{\mathbf{U}}$ has distinct roots

$$\alpha, \alpha^q, \ldots, \alpha^{q^{n-1}}.$$

Moreover, the representation of $\mathbf{S}$ as in Lemma 3.2 is

$$\mathbf{S}_i = Tr_q^{q^n}(\alpha^i) = \sum_{j=0}^{n-1} \alpha^{q^{ji}}.$$

Theorem 3.3. *Suppose* $g(x) = \sum_{k=0}^{p^e-1} a_k x^k$. *For each* k, *let* $k = \sum_{i=0}^{e-1} b_{k,i} p^i$ *where* $0 \le b_{k,i} < p$. *Then the linear complexity of* $\mathbf{S}$ *is*

$$\sum_{a_k \neq 0} \prod_{i=0}^{e-1} \binom{n + b_{k,i} - 1}{b_{k,i}}.$$

Proof. We find $\lambda_p(\mathbf{S})$ and the roots of $f_{\mathbf{S}}$ for $g(x) = x^k$ in four steps.

1. Suppose $g(x) = x^{p^i}$. The minimal length recurrence for $\mathbf{S}$ is the minimal length recurrence for $\mathbf{U}$ with its coefficients raised to the p^ith power. Thus $f_{\mathbf{S}}$ is $f_{\mathbf{U}}$ with its coefficients raised to the p^ith power. Therefore $\lambda_p(\mathbf{S}) = n$ and $f_{\mathbf{S}}$ has roots

$$\alpha^{p^i}, \alpha^{p^i q}, \ldots, \alpha^{p^i q^{n-1}}.$$

2. Suppose $g(x) = x^{bp^i}$, where $1 \le b < p$. Then $\mathbf{S}$ is a b-fold product of the sequence in the preceding case with itself. By Proposition 3.1(2), the set of roots of $f_{\mathbf{S}}$ is a subset of the set of b-fold products of elements $\alpha^{p^i q^j}$, $0 \le j < n$. Not all root products are distinct, so we cannot simply apply Proposition 3.1(2) to get $\lambda_p(\mathbf{S})$ precisely. Nonetheless,

$$\mathbf{S}_k = \left(\sum_{j=0}^{n-1} \alpha^{kp^i q^j} \right)^b \tag{2}$$

$$= \sum_{\bar{r}} \frac{b!}{r_0!\, r_1! \cdots r_{n-1}!} \alpha^{kp^i \sum_{j=0}^{n-1} r_j q^j}, \tag{3}$$

where the sum is taken over $\{\bar{r} = (r_0, r_1, \ldots, r_{n-1})\colon r_j \ge 0 \text{ and } \sum_{j=0}^{n-1} r_j = b\}$. The exponents of α are distinct, and the coefficients of the powers of α are nonzero since $1 \le b < p$. By Lemma 3.2, the roots of $f_{\mathbf{S}}$ are the powers of α appearing in (3). Thus

$$\lambda_p(\mathbf{S}) = \binom{n + b - 1}{b}$$

and $f_{\mathbf{S}}$ has roots

$$\left\{ \alpha^{p^i m}\colon \text{where } m = \sum_{j=0}^{n-1} r_j q^j, \sum_{j=0}^{n-1} r_j = b, \text{ and}, \forall j\colon r_j \ge 0 \right\}.$$

3. Suppose $g(x) = x^k$, $k = \sum_{i=0}^{e-1} p^i b_i$, where $0 \le b_i < p$. $\mathbf{S}$ is an e-fold product of sequences in the preceding case, with distinct i's. The roots of $f_{\mathbf{S}}$ are the distinct products of the roots of the connection polynomials of these sequences, one from each connection polynomial. All root products are distinct by the uniqueness of base p representations and the fact that $p^i < q$ when $i < e$. Thus, by Proposition 3.1(2),

$$\lambda_p(\mathbf{S}) = \prod_{i=0}^{e-1} \binom{n + b_i - 1}{b_i},$$

and $f_{\mathbf{S}}$ has roots

$$\{\alpha^{\sum_{i=0}^{e-1} p^i \sum_{j=0}^{n-1} r_{i,j} q^j}\colon \sum r_{i,j} = b_i \text{ and}, \forall j\colon r_{i,j} \ge 0\}.$$

4. In the general case, S is a sum of sequences of the form considered in the previous case. Due to the uniqueness of base p representations, and the constraints on $\{r_{i,j}\}$, all roots of all the connection polynomials of the summands are distinct. The theorem follows from Proposition 3.1(1). □

Corollary 3.4. *If, for all $x \in GF(q)$, $g(x) \in GF(p)$, then*

$$\lambda_p(\mathbf{S}) = \sum_{a_k \neq 0} \prod_{i=0}^{e-1} \binom{n + b_{k,i} - 1}{b_{k,i}}.$$

Proof. In general, if S is a sequence over a field F, and F is a subfield of E, then the linear complexity of S over F equals its linear complexity over E. To see this, observe that any linear recurrence over F is also a linear recurrence over E, so the linear complexity over E is at most the linear complexity over F. On the other hand, there is a basis for E over F containing 1. Any linear recurrence over E produces one over F by expressing the coefficients in such a basis and considering the component of 1 in all expressions. □

Corollary 3.5. *The linear complexity of* S *(over $GF(q)$ or $GF(p)$) is at most*

$$\binom{n+p-1}{p-1}^{e}.$$

Proof. The maximum occurs when all a_k are nonzero. In that case the linear complexity is

$$\begin{aligned}\sum_{k=0}^{p^e-1} \prod_{i=0}^{e-1} \binom{n + b_{k,i} - 1}{b_{k,i}} &= \sum_{b_0=0}^{p-1} \cdots \sum_{b_{e-1}=0}^{p-1} \prod_{i=0}^{e-1} \binom{n + b_i - 1}{b_i} \\ &= \left(\sum_{b=0}^{p-1} \binom{n+b-1}{b}\right)^{e} \\ &= \binom{n+p-1}{p-1}^{e}.\end{aligned}$$

To see the last equality, we show by induction on t that, for any t and n,

$$\sum_{b=0}^{t} \binom{n+b-1}{b} = \binom{n+t}{t}. \tag{4}$$

When $t = 0$ both sides equal 1. Suppose (4) is true for t replaced by $t - 1$. Then

$$\begin{aligned}\sum_{b=0}^{t} \binom{n+b-1}{b} &= \sum_{b=0}^{t-1} \binom{n+b-1}{b} + \binom{n+t-1}{t} \\ &= \binom{n+t-1}{t-1} + \binom{n+t-1}{t} \\ &= \binom{n+t}{t},\end{aligned}$$

which proves the corollary. □

For a given period $p^r - 1$, we can choose any factorization of r, $r = en$, to produce a geometric sequence of that period. This sequence will be based on an m-sequence of span n over $GF(q)$, $q = p^e$. A natural question is which factorization maximizes the linear complexity, answered by the following proposition.

Proposition 3.6. *If $ne = mf$, $n \neq m$, and n divides m (so f divides e), then, for any t,*

$$\binom{m+t}{t}^f < \binom{n+t}{t}^e.$$

Proof. We may divide e by f and thus assume $f = 1, m = ne$. For every i we have

$$\begin{aligned}(n+i)^e &= i^e + ei^{e-1}n + \cdots \\ &> i^e + ei^{e-1}n \\ &= (ne+i)i^{e-1}.\end{aligned}$$

Therefore

$$\prod_{i=1}^{t} (n+i)^e > \prod_{i=1}^{t} (ne+i)(t!)^{e-1},$$

which implies

$$\begin{aligned}\binom{n+t}{t}^e &= \frac{\prod_{i=1}^{t}(n+i)^e}{(t!)^e} \\ &> \frac{\prod_{i=1}^{t}(ne+i)}{t!} \\ &= \binom{ne+t}{t}.\end{aligned}$$

□

It follows that $\lambda_p(\mathrm{S})$ is maximized by taking $e = r$ and $n = 1$. Unfortunately, this means that the feedforward function g completely defines the sequences, and is defined on an enormous set (of size equal to the period). We have done away with the m-sequence and have an arbitrary binary sequence of period $p^r - 1$.

A more typical use of geometric sequences is to divide the computational work evenly between g and the m-sequence by choosing n approximately equal to $q = p^e$. Then we can generate a sequence of period $q^p - 1$ with a relatively easily defined feedforward function on q elements and a q stage linear feedback shift register over $GF(q)$. We must take care, however, that $g(x)$ cannot be written as $g(x) = h(Tr^{p^e}_{p^f}(x))$ for some f dividing e and h: $GF(p^e) \to \{0, 1\}$. With relatively small q we can obtain sequences of enormous period that are easily generated. The maximum linear complexity we can achieve with such a sequence is

$$\lambda_p(\mathrm{S}) = \binom{p^e + p - 1}{p - 1}^e.$$

This expression will be largest if e is small. Thus choosing q to be a larger prime rather than a power of a small prime will give a larger linear complexity. For example, with $p = 3$ and $q = 27$, we will have a sequence of period approximately

1.4×2^{128}, while the largest linear complexity we can achieve is less than 2^{26}. For $p = 5$ and $q = 25$, the period is approximately 1.1×2^{116}, while the linear complexity is less than 1.1×2^{29}. On the other hand, if $p = q = 17$, the period is only 1.4×2^{69}, but the linear complexity can be made greater than 2^{30}. Even for the latter sequence, however, if bits are generated at the rate of one bit per microsecond, then enough bits to crack the sequence are generated in approximately 1 hour.

We might just as well map $\{0, 1\}$ to an arbitrary pair of elements of $GF(p)$ to consider binary sequences as sequences over $GF(p)$. Let $\sigma: \{0, 1\} \rightarrow GF(p)$ be an arbitrary one-to-one function. We denote by $\sigma(\mathrm{S})$ the result of applying σ to each term of the binary sequence S. By observing that, for any $a \neq b \in GF(p)$, there is a linear transformation of $GF(p)$ mapping $(0, 1)$ onto (a, b), we see

Proposition 3.7. *Let* $\sigma: \{0, 1\} \rightarrow GF(p)$. *Then* $|\lambda_p(\sigma(\mathrm{S})) - \lambda_p(\mathrm{S})| \leq 1$.

4. Finding *p*

The fact that $\lambda_p(\mathrm{S})$ is low is of no use to a cryptanalyst unless p is known. In this section we describe an algorithm that determines p with high probability. We actually determine q with high probability. For purposes of cryptanalysis, however, we only need p.

It was recently shown that, for a geometric sequence S based on an m-sequence over $GF(q)$, q can be determined with high probability if at least $2q^8$ bits of S are known [9]. This attack is based on the calculation of a partial period autocorrelation function. More available bits give a higher probability of success.

We show that, in fact, the same information can be obtained with far fewer bits of the sequence available—approximately q^4 bits—using a simpler statistical measure, the partial imbalance (the imbalance of a sequence is the difference between the number of zeros and the number of ones).

An attack on a system using geometric sequences proceeds in two stages. First, use the partial imbalance to determine q (with high probability). Knowing q tells us p, and this then allows us to use the Berlekamp–Massey algorithm to synthesize a linear feedback shift register over $GF(p)$ that generates S. The number of bits needed for this last stage to work is determined by the linear complexity calculation in Section 3.

Suppose a small partial sequence of S of length D is known. We show that the imbalance of the partial sequence is close to D/q when q is odd. More specifically, we show that, for odd q, the expected imbalance (letting the starting point of the partial sequence vary and keeping the size of the partial sequence fixed) is approximately D/q, and that the variance is sufficiently small that the partial imbalance is close to its expectation with high probability (this is a consequence of Chebyshev's inequality [1]). One way to view our results is that we have introduced a new statistical test that a sequence must satisfy in order to be secure—the variance of the partial imbalance must be high for small partial sequences if the imbalance is high. To simplify notation, we sometimes transform sequences of 0's and 1's into sequences of $+1$'s and -1's. In the case of geometric sequences, we write $G(x) = (-1)^{g(x)}$ to accomplish this.

4.1. *The Imbalance and Partial Imbalance*

Definition 4.1.

1. The imbalance of a binary sequence S of period N is

$$\begin{aligned}\mathcal{I}_{\mathbf{S}} &= \sum_{i=1}^{N} (-1)^{\mathbf{S}_i} \\ &= |\{i\colon \mathbf{S}_i = 0\}| - |\{i\colon \mathbf{S}_i = 1\}| \\ &= 2|\{i\colon \mathbf{S}_i = 0\}| - N.\end{aligned}$$

2. The imbalance of a function $g\colon GF(q) \to GF(2)$ is

$$\begin{aligned}\mathcal{I}_g &= \sum_{x \in GF(q)} G(x) \\ &= |\{x\colon g(x) = 0\}| - |\{x\colon g(x) = 1\}| \\ &= 2|\{x\colon g(x) = 0\}| - q.\end{aligned}$$

Thus we have

Proposition 4.2. *The imbalance of a geometric sequence* S *of period* $q^n - 1$ *with feedforward function* $g\colon GF(q) \to GF(2)$ *is*

$$\mathcal{I}_{\mathbf{S}} = q^{n-1}\mathcal{I}_g - G(0).$$

Proof. This follows from the fact that each element of $GF(q)$ is the image under $Tr_q^{q^n}$ of exactly q^{n-1} elements of $GF(q^n)$, and that $Tr_q^{q^n}(0) = 0$. □

If q is even, then the feedforward function g can be chosen so that $\mathcal{I}_g = 0$. It follows that $\mathcal{I}_{\mathbf{S}} = \pm 1$, and we can learn nothing about S by computing its imbalance. If, however, q is odd, then the smallest imbalance of g that we can achieve is ± 1. In this case we have

$$\mathcal{I}_{\mathbf{S}} = \pm q^{n-1} \pm 1.$$

Thus for such a sequence we can hope to determine q^{n-1} if we can compute $\mathcal{I}_{\mathbf{S}}$. Such a computation is, unfortunately, hopeless—we would need to know the entire sequence S to compute $\mathcal{I}_{\mathbf{S}}$. We can ask, however, how much can be learned about a sequence if only a partial sequence is known. This leads to the following definition.

Definition 4.3. The partial imbalance of a sequence is defined by limiting the range of values in the sum defining the imbalance to a fixed window. It is parametrized by the start position k and length D of the window:

$$\mathcal{I}_{\mathbf{S}}(k, D) = \sum_{i=k}^{D+k-1} (-1)^{\mathbf{S}_i}.$$

If an adversary to a stream cipher system knows D consecutive bits of a sequence, then she can compute a partial imbalance with window size D. The start position k will be unknown to the adversary. If the partial imbalance is sufficiently well behaved for small enough windows and varying k, then the adversary can get useful information. It is hopeless to expect a precise expression for the partial imbalance. We show, however, that the expected partial imbalance (averaged over the starting position of the window) is closely related to the full imbalance. We also show that for certain window sizes the variance of the partial imbalance (with fixed window size D, but varying start position k) is low enough that an adversary has high probability of discovering q. This is a consequence of Chebyshev's inequality [1] which implies that a bound on the variance implies a bound on the probability that a particular partial imbalance is far from the expected partial imbalance.

We begin by showing that the expected partial imbalance of any sequence can be determined from its full imbalance. We denote the expectation of a random variable X by $E[X]$. All expectations are taken for fixed window size D, assuming a uniform distribution on all start positions k.

Theorem 4.4. *Let* S *be a periodic binary sequences with period N. Then the expectation of the partial imbalance of* S *is given by*

$$E[\mathscr{I}_{\mathrm{S}}(k, D)] = \frac{D}{N}\mathscr{I}_{\mathrm{S}}.$$

Proof. Follows by an interchange of summations and a shift of indices to make the double sums independent. □

In the case of interest, we have

Corollary 4.5. *Suppose* S *is a geometric sequence based on an m-sequence of span n with elements in $GF(q)$, with feedforward function g: $GF(q) \to GF(2)$. Assume that* S *is as balanced as possible, i.e., $|\mathscr{I}_g| \le 1$. Then the expected partial imbalance of* S *is*

$$E[\mathscr{I}_{\mathrm{S}}(k, D)] = \frac{D(\pm q^{n-1} \pm 1)}{q^n - 1}$$

if q is odd, and

$$E[\mathscr{I}_{\mathrm{S}}(k, D)] = \pm\frac{D}{q^n - 1}$$

if q is even.

4.2. *The Algorithm*

If q is even, then the expected partial imbalance is too small to hope to recover useful information. If q is odd, then the expected partial imbalance is approximately D/q, and we can hope to learn q from it. If the partial imbalances (for varying k) lie close to the expected partial imbalance, then we can compute a partial imbalance (with q and k unknown), determine which expected partial imbalance $D(\pm q^{n-1} \pm 1)/(q^n - 1)$ the result lies closest to, and conclude that q was used to generate the sequence. That is, we use the following algorithm:

Algorithm for Finding q.

1. **Establish** disjoint intervals U_3, U_5, ... in the positive real line (one for each power of an odd prime, or, for simplicity, one for each odd integer).
2. **Input** a partial sequence $\mathbf{S}_k, \mathbf{S}_{k+1}, \ldots, \mathbf{S}_{k+D-1}$ of $\mathbf{S}$. (Determined, say, by a known plaintext attack on a stream cipher system. The value k is unknown.)
3. **Compute** $x = \sum_{i=k}^{k+D-1}(-1)^{\mathbf{S}_i}$ $(= \mathscr{I}_{\mathbf{S}}(k, D))$.
4. **Find** q such that $|x| \in U_q$.
5. **Output** q.

Is this algorithm likely to be successful? Only if we can choose disjoint intervals U_q so that $\mathscr{I}_{\mathbf{S}}(k, D)$ is in U_q with high probability whenever q was used as the parameter for generating $\mathbf{S}$. If we can sow that the partial imbalances do not deviate too much from their expectations, then it will be possible to choose the intervals U_q. Chebyshev's inequality bounds the deviation of a random variable from its expectation in terms of its variance.

Proposition 4.6 (Chebyshev's Inequality [1]). *If X is a random variable with expectation $E[X]$ and variance $V(X)$, then, for any $\varepsilon > 0$,*

$$\text{Prob}\{|X - E[X]| > \varepsilon\} < \frac{V(X)}{\varepsilon^2}.$$

As it turns out, the algorithm needs at least q^4 bits for success, more than are available if $n < 4$, so we assume $n \geq 4$. If q is an odd prime power, then, for $n \geq 4$,

$$\frac{D(q^3 - 1)}{q^4 - 1} \leq |E[\mathscr{I}_{\mathbf{S}}(k, D)]| \leq \frac{D(q^3 + 1)}{q^4 - 1}.$$

The interval U_q should contain all these points. For each odd q, we pick a positive real number ε_q, such that

$$\frac{D(q^3 + 1)}{q^4 - 1} + \varepsilon_q = \frac{D((q - 2)^3 - 1)}{(q - 2)^4 - 1} - \varepsilon_{q-2}.$$

We then let U_q be the interval from $D(q^3 - 1)/(q^4 - 1) - \varepsilon_q$ to $D(q^3 + 1)/(q^4 - 1) + \varepsilon_q$. If we can choose the ε_q so that

$$\frac{V(\mathscr{I}_{\mathbf{S}}(k, D))}{\varepsilon_q^2} < \tfrac{1}{2}$$

whenever $\mathbf{S}$ is a geometric sequence of period $q^n - 1$ based on a feedforward function g: $GF(q) \to GF(2)$, and $n \geq 4$, then Chebyshev's inequality can be applied to show that the algorithm is successful with probability at least 1/2. In the next section we show that this is the case if enough bits are available, i.e., if D is large enough. Moreover, both $V(\mathscr{I}_{\mathbf{S}}(k, D))$ and ε_q will be proportional to D, so the more bits that are available, the higher the probability of success.

4.3. *The Variance of the Partial Imbalance*

We next consider the variance of the partial imbalance. This section consists of a proof of the following bound, which implies, for example, that $2q^4$ bits of a geometric sequence with balanced feedforward function and odd q are sufficient to determine q with probability at least 1/2. The probability of success goes up if more bits of S are known. We let $v = (q^n - 1)/(q - 1)$.

Theorem 4.7. *For any τ, if $D \leq v$, then the variance of the partial imbalance of a geometric sequence with window D is bounded above by D.*

Proof. Recall that the variance of a random variable X is defined to be $E[(X - E[X])^2] = E[X^2] - E[X]^2$, so we must determine the second moment $E[\mathscr{I}_{\mathrm{S}}(k, D)^2]$ of the partial imbalance. We can reduce this determination to the determination of the cardinalities of certain sets, as stated in the following proposition. By identifying $GF(q^n)$ with n-dimensional affine space over $GF(q)$, these sets are identified with intersections of hyperplanes. If $s \in GF(q)$, and $A \in GF(q^n)$, then we denote by H_A^s the hyperplane $\{x\colon Tr_q^{q^n}(Ax) = s\}$.

Proposition 4.8. *If* S *is a geometric sequence, then*

$$E[\mathscr{I}_{\mathrm{S}}(k, D)^2] = \frac{1}{q^n - 1} \sum_{i,j=0}^{D-1} \left(\sum_{s,t \in GF(q)} N_{i,j}(s, t) G(s) G(t) - 1 \right),$$

where

$$N_{i,j}(s, t) = |H_{\alpha^i}^s \cap H_{\alpha^j}^t|.$$

Proof.

$$\begin{aligned} E[\mathscr{I}_{\mathrm{S}}(k, D)^2] &= \frac{1}{q^n - 1} \sum_{k=0}^{q^n-2} \mathscr{I}_{\mathrm{S}}(k, D)^2 \\ &= \frac{1}{q^n - 1} \sum_{k=0}^{q^n-2} \left(\sum_{i=k}^{k+D-1} G(\alpha^i) \right)^2 \\ &= \frac{1}{q^n - 1} \sum_{k=0}^{q^n-2} \sum_{i,j=k}^{k+D-1} G(\alpha^i) G(\alpha^j) \\ &= \frac{1}{q^n - 1} \sum_{k=0}^{q^n-2} \sum_{i,j=0}^{D-1} G(\alpha^{i+k}) G(\alpha^{j+k}) \\ &= \frac{1}{q^n - 1} \sum_{i,j=0}^{D-1} \sum_{k=0}^{q^n-2} G(\alpha^{i+k}) G(\alpha^{j+k}) \\ &= \frac{1}{q^n - 1} \sum_{i,j=0}^{D-1} \sum_{\substack{x \neq 0 \in \\ GF(q^n)}} G(\alpha^i x) G(\alpha^j x) \\ &= \frac{1}{q^n - 1} \sum_{i,j=0}^{D-1} \left(\sum_{s,t \in GF(q)} N_{i,j}(s, t) G(s) G(t) - 1 \right). \qquad \square \end{aligned}$$

Thus we must determine the values of $N_{i,j}(s, t)$, which can be q^{n-1}, q^{n-2}, or 0. We next analyze the circumstances under which each of these values occurs.

Proposition 4.9. *For any* $0 \le i, j \le q^n - 2$, *and* $s, t \in GF(q)$:

1. *If* $\alpha^{i-j} \notin GF(q)$, *then* $N_{i,j}(s, t) = q^{n-1}$.
2. *If* $\alpha^{i-j} \in GF(q)$ (*i.e., there is an integer m such that* $i - j = m(q^n - 1)/(q - 1)$) *and* $\alpha^i t = \alpha^j s$, *then* $N_{i,j}(s, t) = q^{n-2}$.
3. *If* $\alpha^{i-j} \in GF(q)$ *and* $\alpha^i t \ne \alpha^j s$, *then* $N_{i,j}(s, t) = 0$.

Proof. Two hyperplanes in $GF(q^n)$ are either (1) parallel, in which case they coincide or their intersection is empty, or (2) in general position, in which case their intersection has cardinality q^{n-2}.

Suppose that $\alpha^{i-j} \in GF(q)$. Then $Tr_q^{q^n}(\alpha^j x) = t$ if and only if

$$Tr_q^{q^n}(\alpha^i x) = Tr_q^{q^n}(\alpha^{i-j}\alpha^j x) = \alpha^{i-j} Tr_q^{q^n}(\alpha^j x) = \alpha^{i-j} t.$$

Thus

$$H_{\alpha^i}^s \cap H_{\alpha^j}^t = H_{\alpha^i}^s \cap H_{\alpha^i}^{\alpha^{i-j}t}.$$

This set is empty if $\alpha^j s \ne \alpha^i t$, otherwise it consists of a single hyperplane. In particular, $H_{\alpha^i}^s$ and $H_{\alpha^j}^t$ are parallel.

Conversely, suppose $H_{\alpha^i}^s$ and $H_{\alpha^j}^t$ are parallel. Any hyperplane H_A^u is parallel to H_A^0—for any fixed $y \in GF(q^n)$ such that $Tr_q^{q^n}(Ay) = u$, H_A^u is the translation of H_A^0 by y. It follows that $H_{\alpha^i}^0$ and $H_{\alpha^j}^0$ are parallel. Since $x = 0$ lies on both hyperplanes, they must coincide. That is, $Tr_q^{q^n}(\alpha^i x) = 0$ if and only if $Tr_q^{q^n}(\alpha^j x) = 0$. This implies that the $GF(q)$-linear functions $Tr_q^{q^n}(\alpha^i x)$ and $Tr_q^{q^n}(\alpha^j x)$ agree on $n - 1$ linearly independent vectors $z_1, \ldots, z_{n-1} \in GF(q^n)$. Let z_n be a point of $GF(q^n)$ which is not in $H_{\alpha^i}^0$, and let $u = Tr_q^{q^n}(\alpha^i z_n)$ and $v = Tr_q^{q^n}(\alpha^j z_n)$. Consider the $GF(q)$-linear function

$$f(x) = \frac{u}{v} Tr_q^{q^n}(\alpha^j x).$$

Then f and $Tr_q^{q^n}(\alpha^i x)$ agree on $z_1, \ldots, z_{n-1}$ (both functions are zero on these points). Moreover,

$$\begin{aligned} f(z_n) &= \frac{u}{v} Tr_q^{q^n}(\alpha^j z_n) \\ &= \frac{u}{v} v \\ &= u \\ &= Tr_q^{q^n}(\alpha^i z_n), \end{aligned}$$

so f and $Tr_q^{q^n}(\alpha^i x)$ agree on a set of n independent vectors in $GF(q^n)$, hence agree everywhere. In other words, for all x,

$$\begin{aligned} Tr_q^{q^n}(\alpha^i x) &= f(x) \\ &= \frac{u}{v} Tr_q^{q^n}(\alpha^j x) \\ &= Tr_q^{q^n}\left(\frac{u}{v}\alpha^j x\right). \end{aligned}$$

It follows, that, for all x,

$$Tr_q^{q^n}\left(\left(\alpha^i - \frac{u}{v}\alpha^j\right)x\right) = 0.$$

This is only possible if $\alpha^{i-j} = u/v \in GF(q)$. □

We return to our computation of the second moment of the partial imbalance of geometric sequences with window size $D \leq v$. This implies that $N_{i,j}(s, t) = q^{n-2}$ unless $i = j$, in which case $N_{i,j}(s, t) = q^{n-1}$ if $s = t$, and $N_{i,j}(s, t) = 0$ otherwise. We have

$$\begin{aligned}
E[\mathscr{I}_S(k, D)^2] &= \frac{1}{q^n - 1}\sum_{i,j=0}^{D-1}\left(\sum_{s,t \in GF(q)} N_{i,j}(s, t)G(s)G(t) - 1\right)\\
&= \frac{1}{q^n - 1}\left(\sum_{0 \leq i \neq j < D}\left(\sum_{s,t \in GF(q)} q^{n-2}G(s)G(t) - 1\right)\right.\\
&\qquad \left. + \sum_{i=0}^{D-1}\left(\sum_{s \in GF(q)} q^{n-1}G(s)^2 - 1\right)\right)\\
&= \frac{1}{q^n - 1}\left(\sum_{0 \leq i \neq j < D}(q^{n-2}\mathscr{I}_g^2 - 1) + \sum_{i=0}^{D-1}(q^n - 1)\right)\\
&= \frac{(D^2 - D)(q^{n-2}\mathscr{I}_g^2 - 1)}{q^n - 1} + D.
\end{aligned}$$

It follows that

$$\begin{aligned}
&V(\mathscr{I}_S(k, D))\\
&\quad = E[\mathscr{I}_S(k, D)^2] - E[\mathscr{I}_S(k, D)]^2\\
&\quad = \frac{(D^2 - D)(q^{n-2}\mathscr{I}_g^2 - 1)}{q^n - 1} + D - \left(\frac{D(q^{n-1}\mathscr{I}_g \pm 1)}{q^n - 1}\right)^2\\
&\quad = D\left(1 - \frac{(q^{n-2}\mathscr{I}_g^2 - 1)}{q^n - 1}\right) + \frac{D^2}{(q^n - 1)^2}((q^n - 1)(q^{n-2}\mathscr{I}_g - 1) - (q^{n-1}\mathscr{I}_g \pm 1)^2)\\
&\quad = D\left(1 - \frac{(q^{n-2}\mathscr{I}_g^2 - 1)}{q^n - 1}\right) - \frac{D^2 q^{n-2}(q \pm \mathscr{I}_g)^2}{(q^n - 1)^2}\\
&\quad < D.
\end{aligned}$$

This proves the theorem on the variance of partial imbalance. □

4.4. *Choosing the Intervals*

In this section we describe how the intervals U_q can be chosen so that the results of the previous section, combined with Chebyshev's inequality, imply that the algorithm for finding q is successful with high probability. We assume q is odd and that g is as balanced as possible (that is, $\mathscr{I}_f = \pm 1$). This makes the sequence as statistically random as possible.

As explained in Section 4.2, we find a positive number ε_q for each odd q. It would suffice to do so for each odd prime power q. However, to avoid the difficulty of recognizing prime powers, we simply do so for each odd number. In any case, in the sequence of odd prime powers there will be consecutive odd numbers—e.g., 25 and 27—so the worst-case complexity cannot be improved by restricting q to be an odd prime power.

Proposition 4.10. *Let*

$$\varepsilon_q = \frac{q^2 - 3}{q^4 - 1}.$$

Let

$$\begin{aligned} U_q &= \left[\frac{q^3 - 1}{q^4 - 1} - \varepsilon_q, \frac{q^3 + 1}{q^4 - 1} + \varepsilon_q\right] \\ &= \left[\frac{q^3 - q^2 + 2}{q^4 - 1}, \frac{q^3 + q^2 - 2}{q^4 - 1}\right]. \end{aligned}$$

Then $\{U_q\}$ *are pairwise disjoint intervals in the real line. If* S *is a geometric sequence based on an m-sequence over* $GF(q)$ *of span at least 4, then the interval of radius* ε_q *centred at* $E[I_{\mathrm{S}}(k, D)]$ *is contained in* U_q.

Proof. The first assertion is straightforward. The second assertion holds because, as was shown in Section 4.2,

$$\frac{q^3 - 1}{q^4 - 1} \le E[\mathcal{I}_{\mathrm{S}}(k, D)] \le \frac{q^3 + 1}{q^4 - 1}$$

for the geometric sequences in question. □

We can now combine this with Chebyshev's inequality and our results on the variance of the partial imbalance to obtain the following theorem.

Theorem 4.11. *Let* $n \ge 4$, *and let* S *be a geometric sequence based on an m-sequence of span n over* $GF(q)$ *and a feedforward function that is as balanced as possible. Then the algorithm for determining q using partial imbalances with a window* $D \le v = (q^n - 1)/(q - 1)$ *will succeed with probability at least*

$$1 - \frac{(q^4 - 1)^2}{D(q^2 - 3)^2}.$$

Proof. The probability that the algorithm is successful is the probability that $\mathcal{I}_{\mathrm{S}}(k, D)$ is in U_q. We have

$$\begin{aligned} \mathrm{Prob}_k\{\mathcal{I}_{\mathrm{S}}(k, D) \in U_q\} &\ge \mathrm{Prob}_k\{|\mathcal{I}_{\mathrm{S}}(k, D) - E[\mathcal{I}_{\mathrm{S}}(k, D)]| < \varepsilon_q\} \\ &\ge 1 - \frac{V_k(\mathcal{I}_{\mathrm{S}}(k, D))}{\varepsilon_q^2} \quad \text{(by Chebyshev's inequality)} \\ &\ge 1 - \frac{D}{\varepsilon_q^2} \quad \text{(by Theorem 4.7)} \\ &\ge 1 - \frac{(q^4 - 1)^2}{D(q^2 - 3)^2} \quad \text{(by the definition of } \varepsilon_q\text{).} \end{aligned}$$

□

If $n = 5$ and $q \leq 5$, or if $n = 4$ this probability will be negative for all D, so Chebyshev's inequality does not tell us whether the attack has a positive probability of determining q. However, we have

Corollary 4.12. *If $n = 5$ and $q \geq 7$, or $n \geq 6$, then using a window of size*

$$D > \frac{(q^4 - 1)^2}{(q^2 - 3)^2}$$

gives a positive probability of successfully determining q.

For example, if $q = 27$, then 535,841 bits suffice to determine q with positive probability. Of course we want to determine q with high probability.

Corollary 4.13. *The probability that the algorithm is successful is greater than δ if we use a window size of*

$$D > \frac{(q^4 - 1)^2}{(1 - \delta)(q^2 - 3)^2}.$$

This is possible if n is large enough that

$$\frac{q^n - 1}{q - 1} > \frac{(q^4 - 1)^2}{(1 - \delta)(q^2 - 3)^2}.$$

For example, if $q = 27$ and $n \geq 6$, then we can determine q with probability at least 1/2 if 1,071,682 bits are known, a relatively small number.

5. Conclusions

Perhaps the most important aspect of this paper is the consideration, for a binary sequence, of linear complexity relative to an odd prime number. We have demonstrated that this linear complexity can be far smaller than the period of the sequence, even when the usual linear complexity is quite large. This can be exploited in a cryptologic attack. The belief that high linear complexity gives a degree of security is fallacious. At the very least, the linear complexity must be high relative to all small primes.

We have shown that geometric sequences based on m-sequences over a finite field $GF(q)$ of odd characteristic p can be cracked if enough bits are known. By finding upper bounds on the linear complexity *relative to p* we show that these sequences are vulnerable to a Berlekamp–Massey type attack. The number of bits required depends on the parameters of the sequence (not simply the period). If geometric sequences of this type continue to be used, this dependence and considerations of efficiency should influence the choice of parameters. It seems that it is best to choose p fairly large and generate a sequence of period $p^p - 1$. This gives an easily generated sequence with linear complexity relative to p as large as possible for geometric sequences with approximately this period.

We have also shown that if p is not known, then it can be discovered with high probability if enough bits are known. The algorithm for determining p exploits the lack of balance in geometric sequences and uses a new statistical measure, the partial imbalance. In general, far fewer bits are required to determine p than are required for the Berlekamp–Massey attack. For example, if we use a geometric sequence S based on an m-sequence of span 17 over $GF(17)$, so $n = p = q = 17$, then $\lambda_{17}(S)$ is approximately 1.1×2^{30}. The period of the sequence is approximately 1.4×2^{69}. Thus with 2.2×2^{30} bits available we can determine p with probability at least $1 - 2^{-14}$, and then determine a linear feedback shift register over $GF(17)$ that outputs S. The drawback is that if $\lambda_{17}(S)$ is close to 2^{30}, then this feedback register will have span close to 2^{30} and will generate the sequence much more slowly than the original device using comparable hardware. It is an interesting question whether the information we have acquired can be used to synthesize a faster device for generating the sequence—such as the original device.

It is, of course, dangerous to rely on linear complexity as a measure of cryptographic security. There are many other statistical tests a sequence must pass—in this paper we have shown that the linear complexity relative to primes other than two must be high and the variance of the partial imbalance must be high if the imbalance is large.

Acknowledgements

The author thanks Mark Goresky for many helpful discussions, and Judy Goldsmith for valuable comments on the manuscript.

References

[1] H. Bauer, *Probability Theory and Elements of Measure Theory*, Holt, Rinehart, and Winston, New York, 1972.

[2] M. Blum and S. Micali, How to generate cryptographically strong sequences of pseudo-random bits, *SIAM J. Comput.*, **13** (1984), 850–864.

[3] L. Brynielsson, On the linear complexity of combined shift registers, *Proceedings of Eurocrypt 1985*, Springer-Verlag, Berlin, 1985, pp. 156–160.

[4] A. H. Chan and R. Games, On the linear span of binary sequences from finite geometries, q odd, *Advances in Cryptology: Proceedings of Crypto 1986*, Springer-Verlag, Berlin, 1987, pp. 405–417.

[5] S. Golomb, *Shift Register Sequences*, Aegean Park Press, Laguna Hills, CA, 1982.

[6] T. Herlestam, On linearization of nonlinear combinations of linear shift register sequences, *Proceedings of IEEE ISIT*, Ithaca, NY, 1977.

[7] T. Herlestam, On functions of linear shift register sequences, *Proceedings of Eurocrypt 1985*, Springer-Verlag, Berlin, 1985, pp. 119–129.

[8] A. Klapper, A. H. Chan, and M. Goresky, Cross-correlations of linearly and quadratically related geometric sequences and GMW sequences, *Discrete Appl. Math.*, to appear.

[9] A. Klapper and M. Goresky, Revealing information with partial period autocorrelations, *Proceedings of Asiacrypt '91*, Fujyoshida, Japan.

[10] R. Lidl and H. Niederreiter, *Finite Fields*, Encyclopedia of Mathematics, Vol. 20, Cambridge University Press, Cambridge, 1983.

[11] F. J. MacWilliams and H. B. Hann, On the p-rank of the design matrix of difference set, *Inform. Control*, **12** (1968), 474–488.

[12] J. L. Massey, Shift register sequences and BCH decoding, *IEEE Trans. Inform. Theory*, **15** (1969), 122–127.

[13] R. McElience, *Finite Fields for Computer Scientists and Engineers*, Kluwer Academic, Boston, 1987.

[14] M. Simon, J. Omura, R. Scholtz, and B. Levitt, *Spread-Spectrum Communications*, Vol. 1, Computer Science Press, Rockville, MD, 1985.

[15] K. J. C. Smith, On the p-rank of the incidence matrix of points and hyperplanes in a finite projective geometry, *J. Combin. Theory*, **7** (1969), 122–129.

[16] D. Welsh, *Codes and Cryptography*, Clarendon Press, Oxford, 1988.

[17] A. Yao, Theory and applications of trapdoor functions, *Proceedings of the 23rd IEEE Symposium on Foundations of Computer Science*, 1982, pp. 80–91.

[18] N. Zierler and W. Mills, Products of linearly recurring sequences, *J. Algebra*, **27** (1973), 147–157.